Study of Natural Scene Categories in Measurement of Perceived Image Quality

Xiaohan Yang, Fan Li, Leida Li, Ke Gu, and Hantao Liu

Abstract—**One challenge facing image quality assessment (IQA) is that current models designed or trained on the basis of exiting databases are intrinsically suboptimal and cannot deal with the real-world complexity and diversity of natural scenes. IQA models and databases are heavily skewed towards the visibility of distortions. It is critical to understand the wider determinants of perceived quality and use the new understanding to improve the predictive power of IQA models. Human behavioural categorisation performance is powerful and essential for visual tasks. However, little is known about the impact of natural scene categories on perceived image quality. We hypothesize that different classes of natural scenes influence image quality perception – how image quality is perceived is not only affected by the lower-level image statistics and image structures shared between different categories, but also by the semantic distinctions between these categories. In this paper, we first design and conduct a fully controlled psychovisual experiment to verify our hypothesis. Then, we propose a computational framework that integrates the natural scene category-specific component into image quality prediction. Research demonstrates the importance and plausibility of considering natural scene categories in future IQA databases and models.**

Index Terms—**Image quality, natural scene categories, psychovisual experiment, perception, objective metric**

I. Introduction

NOWADAYS, digital images are widely used in many important research and commercial applications [1]-[4]. However, images are inevitably subject to a variety of distortions in the process of acquisition, compression, transmission, and storage. These distortions result in a degradation in image quality, which affects human's visual experiences. Therefore, it is essential to develop reliable image quality assessment (IQA) methods to quantify image quality in a broad range of applications including image processing, computer vision and pattern recognition [5]-[8].

"Since human beings are the ultimate receivers in most image-processing applications, the most reliable way of assessing the quality of an image is by subjective evaluation. Indeed, the mean opinion score (MOS), a subjective quality measure requiring the services of a number of human observers, has been long regarded as the best method of image quality measurement." [9] Subjective testing must be thoroughly designed and test conditions must be closely controlled so that the variable being measured is statistically meaningful [10]. To guarantee the reliability and statistical significance of the MOS of image quality measurement, subjective experimental protocols and procedures have been developed and adopted as parts of an international standard by the International Telecommunications Union (ITU) [10]. For example, the standard recommends the minimum number of human subjects for a typical subjective image quality assessment experiment, as well as essential experimental settings and subjective data processing steps, etc. Researchers have used the best practice guidance to conduct subjective experiments and generate MOS-based databases that can faithfully reflect human perception of image quality [11]-[12] However, subjective testing is cumbersome, expensive and time-consuming [13], and thus can hardly be used in practical applications. Therefore, a more realistic solution is to develop objective IQA methods that can automatically evaluate image quality as perceived by human beings.

There has been growing interest in developing objective IQA methods. Depending on the availability of the pristine reference image, objective IQA methods are classified into three categories: full-reference IQA (FR-IQA) [14]-[15], reduced-reference IQA (RR-IQA) [16]-[17] and no-reference IQA (NR-IQA) methods. Nevertheless, in many real-world applications, the pristine reference image is often unavailable, which makes FR-IQA and RR-IQA inapplicable. Thus, it has become increasingly important to develop effective NR-IQA methods which can predict image quality without any reference. NR-IQA is a very challenging scientific problem mainly because little is known about the mechanisms of the human visual system (HVS) in determining image quality. In general, existing NR-IQA methods rely on the assumption that the image distortion is the dominant factor for image quality perception [18]-[21]. NR-IQA models have been developed to establish the relationship between image distortions and perceived quality. The traditional approach taken in NR-IQA models is based on extracting image features that explicitly describe distortions, and learning a shallow regression model to map the image representations onto scalar quality scores. In the literature, a majority of NR-IQA models make use

Xiaohan Yang and Fan Li are with Shaanxi Key Laboratory of Deep Space Exploration Intelligent Information Technology, School of Information and Communications Engineering, Xi' an Jiaotong University, Xi'an, 710049, China. (e-mail: yangxiaohan@stu.xjtu.edu.cn; lifan@mail.xjtu.edu.cn).

Leida Li is with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China. (e-mail: ldli@xidian.edu.cn).

Ke Gu is with Faculty of Information Technology, Beijing University of Technology, Engineering Research Center of Intelligent Perception and Autonomous Control, Ministry of Education, Beijing Laboratory of Smart Environmental Protection, Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing Artificial Intelligence Institute, Beijing, 100124, China (e-mail:guke.doctor@gmail.com).

Hantao Liu is with the School of Computer Science and Informatics, Cardiff University, Cardiff, CF243AA, U.K. (e-mail: LiuH35@cardiff.ac.uk).

This work was supported in part by the National Natural Science Foundation of China under Grant (62071369, 62076013), in part by the Beijing Natural Science Foundation under Grant JQ21014.

of the natural scene statistics to extract distortion-related features to evaluate image quality, such as Gaussian scale mixture (GSM) model in the wavelet domain [18], Weibull and Generalized Gaussian distribution (GGD) model in the DCT domain [19]-[20], the GGD model in the spatial domain [21]. Although the traditional NR-IQA methods have achieved good prediction performance in certain databases/applications, the obvious limitation is that these handcrafted features may not be powerful enough to adequately represent complex image structures and distortions. Therefore, there is still considerable room for improvement in NR-IQA models.

A. *Related work*

Recently, researchers attempt to apply deep learning in the development of objective NR-IQA models. A deep neural network (DNN) has proven ability to capture discriminative task-relevant features, which can be a promising method for IQA problem. However, DNN models heavily rely on large-scale annotated data, such as the ImageNet dataset [22]. In the area of image quality, creating such "big" IQA databases is practically very challenging. This is because a meaningful subjective quality label, i.e., MOS must be derived from psychophysical experiments under fully controlled conditions – making a large database increases the number of images/labels at the expense of the reliability of the psychophysical data. To exploit deep learning techniques in the context of the nature of IQA databases, different approaches have been attempted. In a so-called patch-based method [23]-[25], an image is divided into patches with the aim to augment the IQA database. The ground-truth quality label for each image patch is approximated using either the corresponding overall quality score or the score calculated by a traditional objective IQA metric. However, the disadvantage is that the assigned patch label does not accurately and faithfully reflect the actual perceived quality, simply because no psychophysical data is gathered at the patch level. This drawback hinders a DNN-based model's performance in predicting image quality. An alternative approach taken in DNN-based models [26]-[28] is to augment the IQA database by simulating extra new images with distortions similar to the current data. In this approach, transfer learning and domain adaptation are often adopted to improve the sample efficiency and boost learning performance. Some researchers also use multi-task learning methods to reinforce the importance of the characteristics of distortions in a DNN-based IQA model. In [29], the "distortion type" sub-network is constructed and added to a DNN to optimize its learning ability for image quality prediction. In [30], a two-stream sub-networks (representing different distortion forms) is designed and integrated to a DNN-based IQA model.

The common strategy of existing NR-IQA methods focuses on establishing a mapping between image distortions and subjective quality measures. However, other important factors that can influence image quality remain largely unexplored [31]. Because of this single-factor focus, the construction of IQA databases, for example, the widely used databases [32]-[35] has been strongly skewed towards the single determinant of image quality – distortion. This might have caused a potential bias in the development of IQA models and complications in their predictive power [36]. The urgent challenge facing the IQA research is that models designed or trained on the basis of existing IQA databases are intrinsically suboptimal and cannot deal with real-world complexity and diversity of natural image space [37]-[39]. It is critical to go beyond the single factor of distortion and understand the wider determinants of perceived image quality, and then use the new understanding to improve the predictive power of IQA models.

B. *Contributions*

Vision literature reveals that humans are extremely proficient at categorising natural scenes, despite subtle distinctions between heterogeneous classes of natural scenes; they can recognise natural scenes with exposures as brief as 100 ms, and with little time to prepare for the categorisation tasks [40]-[43]. Such powerful human behavioural categorisation performance is essential for visual tasks such as navigation or the recognition of objects in their natural environment [43]-[44]. Little is known about how natural scene categories play a role in image quality assessment, and how to integrate this perceptually relevant aspect to objective IQA models.

An earlier attempt has been made in [45] to investigate the impact of scene category in IQA. A JPEG database (158 distorted images) and a Blur database (158 distorted images) were created; each contained three scene categories (i.e., indoor, outdoor natural, and outdoor manmade). Subjective quality scoring experiments were conducted separately for these two databases, where a controlled lab experiment was used for the JPEG database and an uncontrolled crowdsourcing experiment was used for the Blur database. The limitations of this study are: first, the diversity is scene category is rather limited as only three scene categories were used; second, the impact of scene category on perceived image quality cannot be revealed for the cross-distortion scenario, because subjective ratings generated independently for the JPEG and Blur databases cannot be compared due to the psychometric scale mismatch [46]; third, no objective IQA model was proposed, and the study focused on testing the added value of incorporating hand-crafted scene category features to existing IQA metrics using a capacity-limited shallow regression (i.e., LSVR) method [47]-[48]. Another attempt has been made in [49] to include scene category information in aesthetic quality assessment (AQA), which is a different but related area to IQA. AQA focuses on categorising images into aesthetically higher or lower quality (i.e., a "aesthetics classification" task [50]) and IQA focuses on quantifying the image quality preference induced by visual signal distortions (i.e., a "preference regression" task). Although no psychovisual experiment was conducted in this study, a computational approach was proposed to exploit semantic recognition to improve AQA. This work adopted a multi-tasking learning framework, where the network architecture design and optimisation take into account the specific characteristics of the "aesthetics classification" task. This approach can inspire a design towards a computational framework for scene category-aware IQA. To overcome above challenges, our work aims to (1) design and conduct a new and thorough psychovisual experiment to analyse the impact of diverse natural scene categories on perceived quality, and as a result to faithfully reveal the human behavioural responses to image distortions as a function of natural scene categories; and (2) design and build a computational model, considering the specific characteristics of the "preference regression" IQA task in the model's architecture and optimisation.

The DNN-based multi-task learning framework has been recently exploited to improve IQA [51]-[52]. The IQA model in [51] consists of two sub-networks – a distortion type identification network and a quality prediction network – sharing the early layers. The IQA model in [52] consists of two sub-networks – a natural scene statistics (NSS) feature prediction task and a quality prediction task – sharing a CNN feature extractor. Both models exploit a highly distortion-related feature (i.e., distortion type in [51] and distortion characteristics in [52]) prediction task as an auxiliary task to enhance the network's representation ability for the IQA task. Based on the multi-task learning framework, it is critical to investigate the impact of higher-level HVS features on IQA and construct new plausible auxiliary tasks. However, it should be noted that prior to modelling psychovisual study should be in place to provide the grounding as well as HVS data so that the new auxiliary task can faithfully learn the higher-level HVS features in the presence/context of image distortions. In addition, both models in [51]-[52] adopt the shared layer feature strategy in the network architecture design. However, the strategy is rather straightforward without fine-grained analysis on the influence of different shared layer locations on the performance of the IQA model. It is worthwhile to thoroughly analyse the impact of shared layer locations in order to optimise the network performance.

In this paper, we first investigate the impact of an HVS-based determinant – natural scene categories – on perceived image quality via a psychovisual experiment. In the experimental design, the independent (i.e., scene categories) and dependent (i.e., perceived quality) variables are fully controlled to ensure the results are unbiased and statistical meaningful. Building upon our preliminary work [36], current contribution lies in providing further justifications and analyses to verify our hypothesis. This results in a "Scene Category IQA" database that is the first and largest of its kind. Second, substantial contributions have been made in this paper that after gathering psychovisual evidence and data, we build a new computational model to integrate natural scene category-specific information to objective image quality assessment. The model is based on multi-task learning with deep neural networks, which jointly optimise scene-specific component and distortion-specific component for image quality prediction. In modelling, to leverage deep learning with limited data in IQA, we take advantage of transfer learning combined with dedicated optimisation strategies to enhance sample efficiency and maximise the model's learning performance.

The remainder of this paper is organized as follows. Section II illustrates the psychovisual study and data analysis. Section III describes the proposed computation method and the experimental results. Section IV gives a discussion, and Section V concludes the paper.

II. Psychovisual study and analysis

A. Hypothesis

To enhance an IQA metric's ability in handling complex and diverse natural image space, researchers attempt to incorporate the functional mechanisms of the human visual system (HVS) [32]-[35]. Since human's ability to categorise natural scenes has proven significant in perceiving and understanding visual content [40], we hypothesize that different classes of natural scenes influence image quality perception. How image distortions are perceived may not only affected by the image structure and low-level image statistics shared between different categories, but also by the semantic distinctions between these categories. In the literature, there is a paucity of research on the impact of the natural scene categories on image quality assessment. Most image quality perception studies were conducted using a small number of original visual scenes. Also, visual scenes were randomly selected without a systematic way of content classification. This poses difficulties for studying the influence of scene categories on image quality. It should be noted that a perception study must be conducted under fully controlled experimental conditions (with minimum uncontrolled variables in the experimental design), otherwise, the findings cannot faithfully reflect human sensory perception [53]. We recently created a new IQA database including natural scene categories, namely the CUID database as detailed in [36]. We now briefly summarise the database, and give further analysis on natural scene categories.

B. The CUID database

A total of sixty source images (original visual scenes) were collected from the Unsplash website [54]. They were high-quality images and had a resolution of 1920 × 1080 pixels. Ten different natural scene categories (six images were chosen to capture the high variability within each category) were purposely selected in a systematic way including ACT (Action), BNW (Black and White), CGI (Computer-Generated Imagery), IND (Indoor), OBJ (Object), ODM (Outdoor Man-made), ODN (Outdoor Natural), PAT (Pattern), POT (Portrait), and SOC (Social). These ten categories of sixty source images are illustrated in Fig. 1.

The original images were distorted by applying three different types of common image distortions: contrast change (i.e., CC), JPEG compression (i.e., JPEG), and motion blur (i.e., MB). These different distortion types essentially give distinctive impairments in images. By varying the distortion parameters, the strength of distortion is adjusted, which generates distorted images of varying quality. Fig. 2 shows an example of distortion simulations, where a source image leads to nine distorted images. For each distortion type, three different levels of distortion/quality (i.e., Q1, Q2 and Q3) are stimulated, reflecting distinctive levels of perceived quality: Q1 indicates 'perceptible but not annoying artifacts', Q2 indicates 'noticeable and annoying artifacts', and Q3 indicates 'very annoying artifacts'. This results in a total of 600 test stimuli (including the original visual scenes).

A perception experiment was carried out at a laboratory at School of Computer Science and Informatics, Cardiff

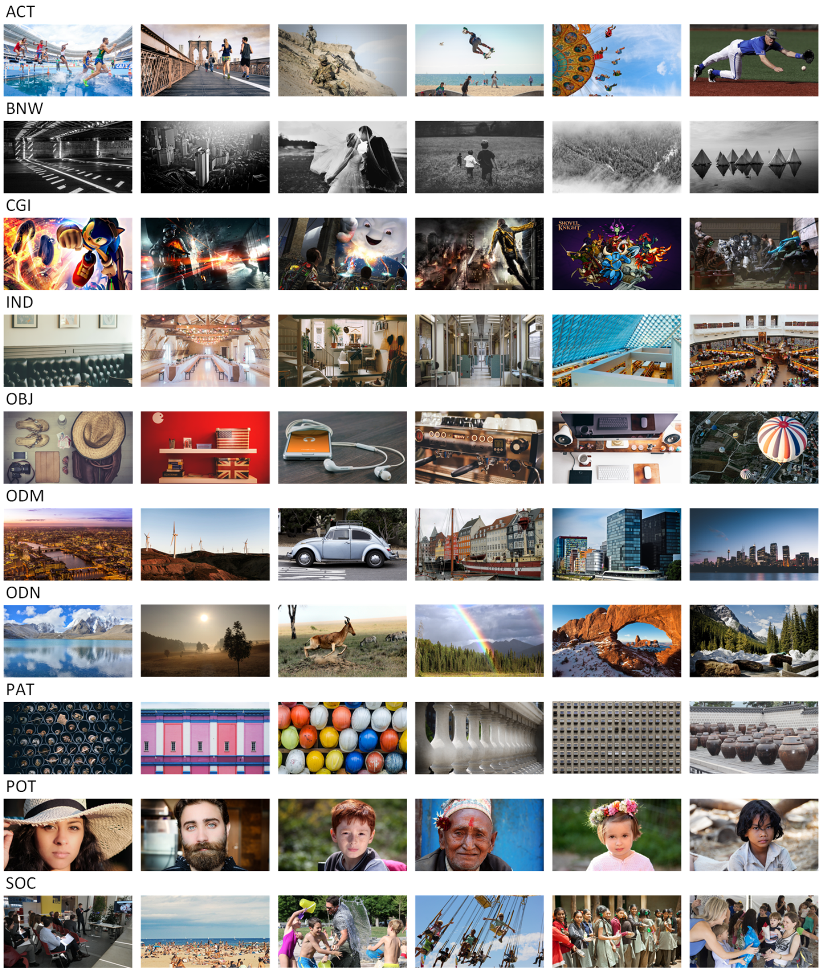


Fig. 1. Ten categories of 60 source images contained in the CUID dataset [36].

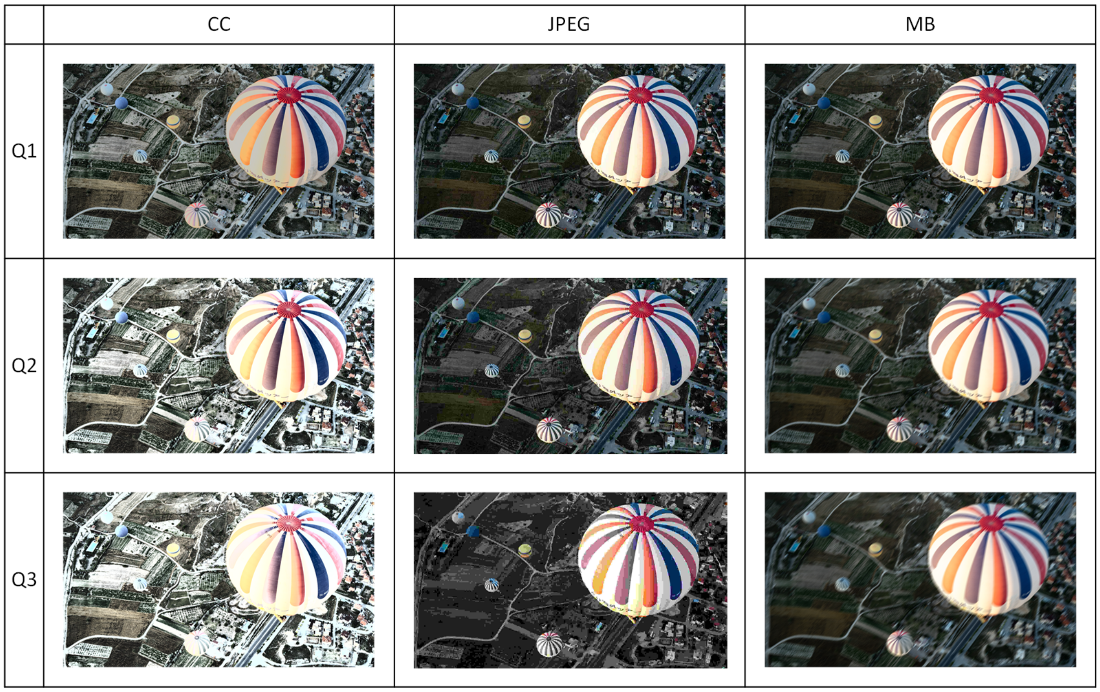


Fig. 2. Exemplars of distorted stimuli contained in the CUID dataset [36].

University. The laboratory was set up as a standard office environment with fully controlled viewing conditions [10]. A 19-inch LCD monitor was used to display the test stimuli. The viewing distance was approximately 60 cm. The experimental procedures followed a single-stimulus method as prescribed by [**?**]. Participants scored image quality using a rating scale that ranged from 0 to 100. The within-subjects experimental design [76] was adopted to ensure subjective results are reliable and consistent. This means each participant must view and score all stimuli in the entire dataset. To eliminate the undesirable carry-over effects due to participant fatigue or boredom [53], multiple sessions were arranged for each participant to complete the rating task as detailed in [36]. Nineteen participants were recruited to take part in the experiments. They were 8 males and 11 females, between age 23-52, and inexperienced with subjective image quality assessment. The characteristics of the assessment panel were determined in accordance with the standard in [10]. In order to make participants familiar with the test stimuli and the use of scoring scale, a training session was provided to each participant before they started the actual rating session.

After stimuli are evaluated by the assessment panel, the mean opinion score (MOS) – representing the overall subjective quality of an image – is derived as the average of individual subjective scores [9]. To account for the potential differences between participants when using the rating scale, z-score is calculated to convert a raw subjective score into a standard (calibrated/normalised) score [55]:

$$ZS_{ij} = (RS_{ij} - \mu_i)/\sigma_i \tag{1}$$

where RS_{ij} indicates the raw-score of the j-th test image rated by the i-th participant, μ_i indicates the mean of all raw-scores given by the participant i, and σ_i indicates the corresponding standard deviation.

A standard procedure to remove outliers (detailed in [55]) was applied. Ultimately, MOS was calculated:

$$MOS = \sum_{i=1}^{P} ZS_{ij} \tag{2}$$

where P denotes the number of scores (excluding outliers) for the j-th image. After MOS values are generated, they are linearly mapped to the range of [0, 100] to match with the original value range of the rating scale. This results in a Cardiff University Image quality Database (CUID). For a well-balanced IQA database, the MOS values of test stimuli should have a uniform distribution across the range of perceived quality. The MOS distribution of the CUID database (see detail in [36]) shows that the test images are, to some extent, evenly distributed across the quality range, which is consistent to other widely recognised IQA databases, such as the LIVE database [32]. The reliability measure of the MOS as per [56] – Pearson correlation between MOS values and individual ratings (IR), i.e. MOSIR is calculated for individual subjects. The 95 % confidence interval of the MOSIR is [0.75, 0.8], indicating a subjective database of high reliability

C. Analysis of MOS and natural scene categories

Now, for the CUID database, the unique new feature is that natural scene categories have been systematically built into the database. Since the within-subjects design was used to generate the MOS, the MOS of an image from one category can be fairly compared to the MOS of an image from any other category [53], without any additional experiments for scale realignment [32].

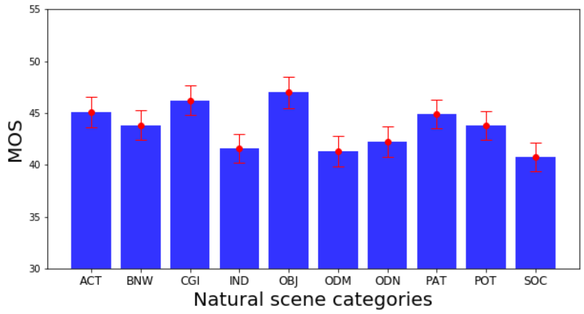


Fig. 3. The CUID database: natural scene ”category-wise” MOS. Error bars indicate a 95% confidence interval.

Fig. 3 shows the natural scene ”category-wise” MOS of the CUID database. As can be seen from the figure, when the same distortions equally were applied to each category of natural scenes in the CUID database, the perceived quality of OBJ category and CGI category is higher than that of other categories. This might suggest OBJ and CGI images are less impacted by the same distortions used in the CUID database. The SOC category is largely affected by distortions, resulting in a lower perceived image quality. The above observation implies that natural scene categories tend to impact perceived image quality. This impact might be attributed to the human cognitive processes, such as emotion or aesthetics. The observed tendencies are further statistically analysed. An analysis of variance (ANOVA) is conducted by selecting perceived quality as the dependent variable, and the categorical natural scene as the independent variable. The ANOVA results show that the categorical natural scene has a statistically significant effect on perceived quality (F-value=8.63, p-value=5.17E-13<0.001 at 95% level).

III. The proposed computational method

We propose a computational framework for the integration of natural Scene Categories in Image Quality prediction, namely SCIQ. The schematic overview of the proposed framework is illustrated in Fig. 4. The framework is based on a multi-task deep neural network, which contains two branches respectively addressing the influence of natural scene categories and distortions. The scene category-specific branch is trained to classify natural scenes, and its output probability of classification is used to guide the quality prediction branch. By doing this, the two branches are jointly optimised to learn the interactions of image distortions and natural scene categories for the image quality prediction task.

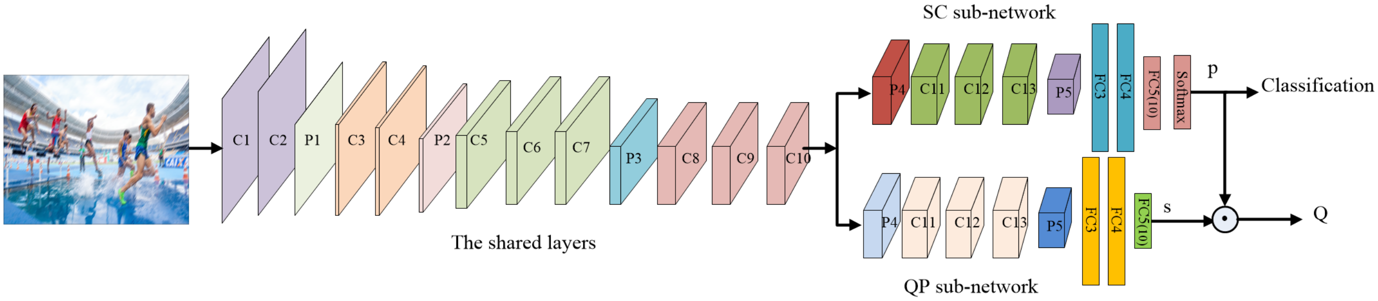


Fig. 4. The proposed framework of SCIQ with a scene category (SC) sub-network and a quality prediction (QP) sub-network.

A. The proposed SCIQ model

1) The SCIQ architecture: The SCIQ architecture contains two sub-networks, including the scene category (SC) sub-network and the quality prediction (QP) sub-network. We adopt the architecture of the pre-trained VGG network [59], including 13 convolution (C) layers, 5 max-pooling (P) layers and 3 fully-connected (FC) layers. The large-scale ImageNet dataset [22] has been used to pre-train the VGG network. By using the pre-trained VGG, the learned parameters/weights of the network can be transferred to the SCIQ task to significantly improve the sample efficiency.

For the scene category (SC) sub-network, since the pre-trained VGG network is powerful for the image classification task, it has potential for the natural scene classification in our model. The number of the last FC layer is modified to n, which aims to discriminate scene categories in the CUID database. Then, the softmax layer is used to obtain the accuracy of scene category, as shown in equation 3.

$$\vec{P} = [p_1, p_1, ..., p_n] \tag{3}$$

where $\vec{P}$ denotes the outputs of the last FC layer in the scene category sub-network, the number of outputs is 10. p_i (i=1 to n) denotes the probability of the ith scene category.

Similarly, the quality prediction (QP) sub-network is also based on the pre-trained VGG network. The number of outputs of the last FC layer is altered to 10, which is to obtain a vector of image quality scores, as show in equation 4.

$$\vec{S} = [s_1, s_1, ..., s_n] \tag{4}$$

where $\vec{S}$ denotes the outputs of the last FC layer in the quality prediction sub-network, the number of outputs is 10. s_i (i=1 to n) denotes the quality score of the ith instance.

The two sub-networks share the mid-level deep features (i.e., at the 10th C layer in our experiment), which aims to speed up the feature discrimination for different tasks [60], [61]. Finally, the image quality Q is obtained by associating the image quality score vector $\vec{S}$ with the corresponding scene category vector $\vec{P}$.

$$Q = \vec{S} \odot \vec{P} = \sum_{i=1}^{n} s_i p_i \tag{5}$$

where $\odot$ represents the weighted sum of element-wise multiplication.

2) The design choices: In our SCIQ algorithm, the mid-level shared features design is proposed, as shown in Fig. 4. This design is based on the fact that the shallow layers of a DNN contain general features, such as edges and textures and deep layers contain specific features, such as higher-level semantics [66]. For multi-task learning, a mid-level shared features design gives a good balance between minimizing training parameters and extracting common attributes for distinctive tasks. The network architecture also features the discrimination ability of specific tasks by providing a transition from common attributes to specific attributes. In principle, the two sub-tasks, scene category (SC) and quality prediction (QP) share some common attributes, such as saliency; while they have their specific attributes, such as local properties for SC and local distortions for QP [28], [61]. So, the mid-level shared features design well captures this property of multi-task learning.

3) The loss function: The two sub-networks are jointly trained using the following loss function L.

$$L = \lambda_1 L_1(w; \theta) + \lambda_2 L_2(w; \theta) \tag{6}$$

where L_1 is the cross entropy loss function [59] of scene category (SC) sub-task. L_2 represents the squared Euclidean distance as the loss function [26] of the quality prediction (QP) sub-task. λ_1 and λ_2 control the two components of the final combined loss function.

4) The training strategy: Prior to training the SCIQ model, the parameters contained in the first 10th convolution layers of the pre-trained VGG-network are shared. The rest parameters are initialized randomly. The last FC layer of the pre-trained VGG network is modified to a ten-dimensional output to suit the scene category and quality prediction sub-tasks.

Then, the SCIQ is trained by using the CUID database. The input image is cropped randomly. The size of cropped images is 224 × 224 pixels. The label for training the scene category sub-network is a vector containing ten elements, indicating the likelihood of scene categories. Meanwhile, the label for training the quality prediction sub-network is a vector of ten quality scores, indicating the ground truth image quality. Finally, the end-to-end optimization strategy is adapted to minimize the losses of the two sub-networks.

B. Experimental results

1) Experimental setup: To evaluate the performance of an image quality metric, two commonly used measures are quantified. They are PLCC (i.e., Pearson Linear Correlation Coefficient) and SROCC (i.e., Spearman Rank-Order Correlation Coefficient) calculated between the estimated visual quality scores Q_{pre} and the subjective quality scores Q_{sub}, as:

$$SROCC(Q_{pre}, Q_{sub}) = 1 - \frac{6\sum d_i}{m(m^2-1)} \tag{7}$$

$$PLCC(Q_{pre}, Q_{sub}) = \frac{cov(Q_{sub}, Q_{pre})}{\sigma(Q_{sub})\sigma(Q_{pre})} \tag{8}$$

where m indicates the number of test stimuli; d_i indicates the rank difference of the i th test sample; $cov(.)$ represents the covariance between Q_{pre} and Q_{sub}; $\sigma(.)$ represents the standard deviation. PLCC measures the prediction accuracy and SROCC measures the prediction monotonicity. The magnitude of both correlation measures ranges from 0 to 1, with 0 indicating no correlation and 1 indicating perfect correlation. Therefore, the larger the measure, the better the model's performance in predicting the subjective image quality [62]-[63].

In training the SCIQ model, we randomly divide the distorted images of each scene category in the CUID database into a training set and a test set. The training set includes four source scenes and the test set includes the two source scenes. By doing so, no overlap occurs between the training set and test set.

The SCIQ model is trained using the Caffe framework. In our experiment, We set the min-batch to be 11; the momentum and weight decay to be 0.9 and 0.0005, respectively; and the learning rate to be 1e-6. Also, we make the training rates decrease by a factor of 0.1 per 10K iterations for a total of 50K iterations. We set the dropout regularization ratio to be 0.5. Note these are commonly used settings for the hyper-parameters of the Caffe deep learning framework [64]-[65].

The relative importance weights are set to be 0.8 and 0.2 in equation 6. The value range of the loss function of the scene category (SC) subtask is between 0 and 1; while the value range of the loss function of the quality prediction (QP) subtask is found to be wider than that of SC subtask. To compensate for the difference between two loss functions and balance their contributions towards the combined loss, we assign a relatively larger weight value to SC subtask and a relatively smaller weight value to QP subtask. To verify the weight assignment and demonstrate how difference assignment combinations can impact the model performance, we conduct experiments and the results are listed in Table I. It shows that the best performance of the SCIQ model is achieved when λ_1=0.8 and λ_2=0.2 are used.

This training process is repeated six times to eliminate the performance bias. For each repetition, the training and test sets are randomly selected as described above. The average values of the SROCC and PLCC are reported as the final results. Note, increasing the times of model running may help reduce possible fluctuation in performance. We run an experiment to increase the times of running the model from six to ten, and compare the model's performance as reported in Table II. It can be seen from the table that the model has reached stable performance when running it for six times.

TABLE I
The performance (i.e., SROCC and PLCC) of the proposed SCIQ model using different λ_1 and λ_2 settings for the combined loss function.

λ_1	λ_2	SROCC	PLCC
0.2	0.8	0.856	0.860
0.4	0.6	0.861	0.866
0.5	0.5	0.878	0.875
0.6	0.4	0.880	0.884
0.9	0.1	0.885	0.890
0.8	0.2	**0.909**	**0.905**

TABLE II
The impact of model running times on the performance (i.e., SROCC and PLCC) of the proposed SCIQ model.

Running of SCIQ model	SROCC	PLCC
Six times	0.909	0.905
Ten times	0.910	0.903

2) Performance on the CUID database: Now, we want to verify the proposed design: (1) whether the SC guidance actually contributes to the prediction power of the network; (2) after which convolution layer the network should break up in to two sub-networks. We run experiments with the CUID database using various design options: (1) a DNN without the SC sub-network; (2) a DNN with both the SC and QP sub-networks; (3) a DNN with the SC and QP sub-network breaking up at different places (i.e., the number of first convolution layers of the pre-trained VGG network used as shared feature layers). As can be seen in Fig. 5 that the prediction power (as measured by the Spearman rank order correlation coefficient (SROCC) between the MOS and predictions) with the scene category guidance is higher than that without the guidance. Also, for the network with the SC guidance, the best performance is achieved when the network breaks up at the 10th convolution layer, meaning the first 10 convolution (C1-C10) layers are used as shared feature layers. This result is in line with the mid-level shared features design in [29], which suggests that the choice should be at the mid-level layers (i.e., C5-C10) to avoid shared features being too general or too specific.

We compare the performance of our proposed SCIQ model to the state-of-the-art image quality assessment (IQA) algorithms, including both Full-reference (FR) and No-reference (NR) models. The FR models include PSNR [67], SSIM [14] and VIF [68]. The NR models include traditional and deep learning-based methods. Since our proposed model is DNN-based, we decided to include only two representative traditional NR-IQA, i.e., BLIINDSSII [20], and BRISQUE [21], and focus on the comparison amongst deep learning-based IQA models. It should be noted that a fair comparison is possible only when the source code is available for all IQA models under study, and the same fine-tuning procedure is consistently applied for all models. We include nine DNN-

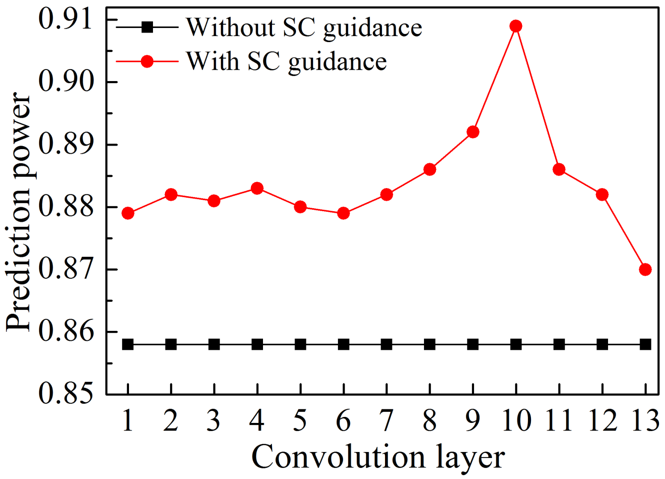


Fig. 5. The prediction power of the network with various design choices.

based NR-IQA, i.e., Alexnet [75], VGG [59], CNN [23], BIECON [69], DIQaM [71], RankIQA[70], WaDIQaM [72], GraphIQA [73], and TTL_SFTnet [74]. Note Alexnet [75] and VGG [59] represent two well-known general DNN models. We simply applied them for the IQA task by first pre-training the model with the ImageNet [22] database and then directly fine-tuning the model on the CUID database. The other models are specifically designed for IQA, therefore, they were directly fine-tuned on the CUID database. Note the same fine-tuning procedure is consistently applied for those nine IQA models to ensure the results are comparable. To the best of our knowledge, this represents a comprehensive comparison of state-of-the-art deep learning-based IQA models that have made their source code publicly available so far in the literature.

Table III shows the performance of these IQA methods. It can be seen that our proposed SCIQ outperforms other IQA metrics.

TABLE III
The performance of different IQA methods on the CUID database. Note for DNN-based NR-IQA, only the IQA models with their source code made publicly available so far in the literature are included in our comparative experiment.

Type	Method	SROCC	PLCC
FR-IQA	PSNR	0.123	0.109
	SSIM	0.522	0.494
	VIF	0.697	0.598
Traditional NR-IQA	BLIINDSSII	0.716	0.729
	BRISQUE	0.722	0.736
DNN-based NR-IQA	AlexNet	0.794	0.797
	VGG	0.858	0.870
	CNN	0.533	0.515
	BIECON	0.772	0.752
	DIQaM	0.847	0.845
	RankIQA	0.867	0.866
	WaDIQaM	0.857	0.853
	GraphIQA	0.803	0.817
	TTL_SFTnet	0.862	0.868
	Our SCIQ	**0.909**	**0.905**

The performance of FR-IQA is unsatisfactory as the correlation (i.e. in terms of PLCC and SROCC) is rather low. For the traditional NR-IQA, the limitation mainly lies in the use of handcrafted features, which can not adequately capture the perceptual characteristics of the combination of image content and distortions, and therefore, their prediction performance is also quite low. Most Deep learning-based methods give good performance due to the fact that deep features representing perceptual image quality can be automatically extracted. Amongst those deep learning-based metrics, Our SCIQ model is the best, possibly due to the scene categories are explicitly included and quantified.

3) Cross-database evaluation: In the literature, cross-database evaluation is often used in the image quality community to measure the generalization ability of IQA models, particularly for machine learning-based and deep learning-based models [24],[27],[70]. The performance of an IQA model could be evaluated on the CUID database only using the conventional train-test split technique as the procedure used in Section III.B 2). However, a more critical performance evaluation would be to use the CUID database as the training set and use a different unseen IQA database (i.e., obtained from a different laboratory, such as the popular LIVE, CSIQ, TID2013 or LIVEMD database [32]-[34]) as the test set. This cross-database evaluation can reveal how well the model can generalise – the ability of a learning model to perform accurately on new, unseen examples after having learned a training set. If a learning model successfully built a general model about the IQA space using the training examples of CUID, it would produce sufficiently accurate predictions for, e.g., LIVE, CSIQ, TID2013 and LIVEMD. As shown in Table IV, compared to other state-of-the-art deep learning-based IQA models, i.e., AlexNet, VGG, CNN, BIECON, DIQaM, RankIQA, WaDIQaM, GraphIQA and TTL_SFTnet, our proposed SCIQ model shows superior generalization ability in the demanding cross-database evaluation.

4) Ablation experiments: A series of systematic ablation experiments are carried out to further verify the rationality of our proposed SCIQ model. Note some ad hoc ablation experiments have been initially conducted in Section III B.2).

Contribution of “shared layer features”: We run four comparative experiments to verify the effectiveness of the mid-level shared features (SF) design choice. In the first experiment (i.e., referred to as “QP-only”), the quality prediction (QP) sub-network is trained and model is rendered as the image quality predictor. In the second experiment (i.e., referred to as “SC-only”), the scene category (SC) sub-network is first trained to classify 10 categories; then the last FC layer of the sub-network is modified to 1 and network is train to predict image quality. In the third experiment (i.e., referred to as “QP-SC-without SF”), the scene category sub-network and quality prediction sub-network are trained separately without considering the mid-level shared features, and the last FC layer is fused to produce a score as the prediction of image quality. In the last experiment (i.e., referred to as “QP-SC-with SF”), the proposed SCIQ architecture is used. The results are listed in Table V. It can be seen that our SCIQ architecture is superior to the method without considering the mid-level shared features.

TABLE IV
Cross-database evaluation. Model performance is quantified by SROCC (note, PLCC exhibits the same trend of SROCC). Note only the DNN-based IQA models with their source code made publicly available so far in the literature are included in our comparative experiment.

Train	Test	AlexNet	VGG	CNN	BIECON	DIQaM	RankIQA	WaDIQaM	GraphIQA	TTL-SFTnet	SCIQ
CUID	LIVE	0.602	0.655	0.528	0.807	0.713	0.752	0.710	0.679	0.761	**0.853**
CUID	CSIQ	0.586	0.602	0.505	0.791	0.695	0.758	0.677	0.658	0.728	0.783
CUID	TID2013	0.597	0.638	0.500	0.560	0.516	0.713	0.702	0.635	0.726	**0.751**
CUID	LIVEMD	0.610	0.643	0.541	0.712	0.661	0.705	0.690	0.600	0.746	**0.773**

TABLE V
The contribution of shared features (SF) to SCIQ design.

Different model design options	SROCC	PLCC
QP-only	0.853	0.870
SC-only	0.860	0.874
QP-SC-without SF	0.866	0.868
QP-SC-with SF (proposed SCIQ)	0.909	0.905

Contribution of “natural scene categories”: To verify the rationality of including the natural scene categories, we run two comparative experiments. In the first experiment (i.e., referred to as “Direct-SC”), without using the scene category (SC) sub-network, we directly provide the SC vector to the network; and use the vector to weight the quality score vector obtained from the last FC layer of the quality prediction (QP) sub-network to generate the final quality score. In the second experiment (i.e., referred to as “Learned-SC”), we use our SCIQ architecture to adaptively learn the relationships between scene categories and image distortions. The results are shown in Table VI. It can be seen that including the scene category (SC) sub-network enhances learning the complex relationships between natural scene categories and image quality assessment.

TABLE VI
The contribution of natural scene categories to SCIQ design.

Different model design options	SROCC	PLCC
Direct-SC	0.879	0.879
Learned-SC (proposed SCIQ)	0.909	0.905

The contribution of “core modelling strategies”: To verify the rationality of our core modelling strategies (i.e., “transfer learning (TL)” and “shared features (SF)”, we run four comparative experiments. In the first experiment (i.e., referred to as “NO TF & NO SF”), the model does not contain shared features between SC and QP sub-networks and is trained directly on the CUID database without transferring information from the pre-trained VGG. In the second experiment (i.e., referred to as “NO TF & YES SF”), shared features design is included, but model is rendered without transfer learning. In the third experiment (i.e., referred to as “YES TF & NO SF”), the model does not include the shared feature design but does make use of transfer learning. In the forth experiment (i.e., referred to as “YES TF & YES SF”), the complete SCIQ design is used including both transfer learning and shared feature design. The results are shown in Table VII. It can be seen that both core modelling strategies significantly contribute to the proposed SCIQ model.

TABLE VII
The contribution of core modelling strategies (i.e., transfer learning (TF) and shared feature (SF)) to SCIQ design.

Different model design options	SROCC	PLCC
NO TF & NO SF	0.523	0.506
NO TF & YES SF	0.569	0.581
YES TF & NO SF	0.866	0.868
YES TF & YES SF (proposed SCIQ)	0.909	0.905

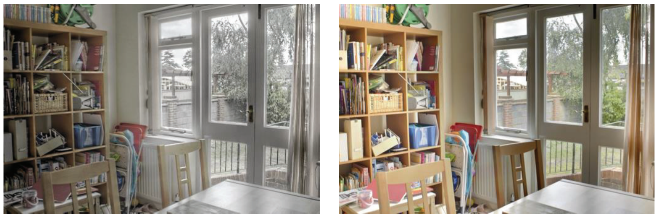
IV. Discussion

In this paper, we focus on single-distortion image quality assessment, where each stimulus is degraded by only one of many possible distortion types. The vast majority of literature has been focusing on single-distortion IQA mainly because the impact of individual distortion types on perceived quality can be thoroughly studied. It should be noted in practical imaging chain, the images often undergo multiple stages of quality degradation, therefore, multiple-distortion image quality assessment is of high practical relevance [34],[78]. An immediate extension of research is to build upon the methodologies established in the current work and investigate the ”natural scene categories” in multiple-distortion IQA by simulating multiple distortion stages and generating multiply distorted images.

To facilitate image quality research, more psychovisual studies should be conducted to provide a better understanding of wider determinants of image quality as perceived by human beings. However, it should be noted that subjective data are meaningless unless they are gathered by well-designed psychometric tests with fully controlled experimental conditions. Also, a great deal of attention has been paid to the image quality assessment behaviour of an average human observer (i.e., MOS), but little attention has been paid to the subjectivity of individuals. Future work could investigate the variances between subjective opinions and their implications on objective IQA models.

To facilitate the development of advanced DNN-based IQA, it is important for developers to make the source code of IQA models as well as IQA databases publicly available so that a fair comparative study can be conducted. Table III so far represents a comprehensive comparison of DNN-based IQA models that have made their source code publicly available. Further comparison can be easily conducted if new IQA source code is released in future.

There is a growing trend to use image quality methodologies to advance technology developments in emerging applications, such as video blending, underwater image enhancement, and image fusion, etc. [79]-[80]. One way to develop useful application-specific IQA models is to understand the characteristics of the visual stimuli through subjective evaluation, and build these application-specific features into objective IQA models. The latter involves an important step to test whether existing IQA models are readily applicable, if so, these models could be modified or adapted to the new application domain. Here, we give an exploratory example of using the proposed SCIQ model in image fusion. Fig. 6 shows two fused images created by different image fusion methods, namely G12 [81] and ShutaoLi12 [82]; and their ground truth image quality (i.e., MOS) scores derived from subjective experiments [83]. Our proposed SCIQ is directly applied to produce objective scores for the fused images, as the results shown in Fig. 6. It can be seen that there is a good level of agreement between the subjective and objective scores, suggesting that our proposed SCIQ model has the potential to be used for assessing the output quality of image fusion methods. However, to effectively adapt our SCIQ model to image fusion, more work is needed including gathering reliable subjective data – output quality of various image fusion methods – via psychovisual experiments as per methodologies used in [81]-[83], and fine turning the SCIQ model on the new subjective image quality scores.



(a) G12 (MOS=3.957 vs. SCIQ=2.798)

(b) ShutaoLi12 (MOS=8.739 vs. SCIQ=8.024)

Fig. 6. An example of two fused images created by different image fusion methods. (a) Fused image generated by G12 [81]. (b) Fused image generated by ShutaoLi12 [82]. Ground truth image quality (i.e., mean opinion score – MOS, note the range is [1,10]) derived from subjective experiments [83] versus objective image quality predicted by our proposed SCIQ model (note, SCIQ is directly applied without fine-tuning on the fused images) is illustrated for each image.

V. Conclusion

In this paper, we have verified an important hypothesis that natural scene categories significantly impact image quality assessment. Through the design and conduct of a fully controlled psychovisual experiment, we found that when the same distortions are applied, different categories of natural scenes intrinsically induce different human behavioural responses to image quality. Building on this psychovisual evidence, we have proposed a computational framework that integrates the natural scene category-specific component to image quality prediction. We have demonstrated the importance of natural scene categories in improving the reliability of image quality models. We suggest that future research should consider natural scene categories in both subjective and objective image quality assessment.

References

[1] F. Li, S. Fu, Z. Li and X. Qian, “A cost-constrained video quality satisfaction study on mobile device,” *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1154–1168, 2018.

[2] Q. Jiang, W. zhou, X. Chai, G. Yue, F. Shao and Z. Chen, “A full-reference stereoscopic image quality measurement via hierarchical deep feature degradation fusion,” *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 12, pp. 9784-9796, 2020.

[3] Q. Jiang, W. Gao, S. Wang, G. Yue, F. Shao, Y. Ho and S. Kwong, “Blind image quality measurement by exploiting high-order statistics with deep dictionary encoding network,” *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7398-7410, 2020.

[4] A. Angelis, A. Moschitta, F. Russo and P. Carbone, “A vector approach for image quality assessment and some metrological considerations,” *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 1, pp. 14-25, 2009.

[5] X. Yang, F. Li, and H. Liu, “A comparative study of DNN-based models for blind image quality prediction,” *in Proc. ICIP*, pp.1019-1023, 2019.

[6] G. Yue, C. Hou, T. Zhou and X. Zhang, “Effective and efficient blind quality evaluator for contrast distorted images,” *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 8, pp. 2733-2741, 2019.

[7] H. Sellahewa and S. Jassim, “Image-quality-based adaptive face recognition,” *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 805-813, 2010.

[8] X. Luo, J. Zhang and Q. Dai, “Saliency-based geometry measurement for image fusion performance,” *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 4, pp. 1130-1132, 2012.

[9] Zhou Wang, Alan C. Bovik, *Modern Image Quality Assessment*, Morgan and Claypool, 2006.

[10] Recommendation ITU-R, BT.500-11, *Methodology for the subjective assessment of the quality of television pictures*, 2002.

[11] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, “Image database TID2013: Peculiarities, results and perspectives,” *Signal Process., Image Commun.*, vol. 30, pp. 57-77, 2015.

[12] P. Le Callet and F. Autrusseau, “Subjective quality assessment IRCCyN/IVC database,” 2005 [Online]. Available: http://www.irccyn.ecnantes.fr/ivcdb/

[13] Z. Wang and A. C. Bovik, “Reduced- and no-reference image quality assessment,” *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 29-40, 2011.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600-612, 2004.

[15] W. Kim, A. Nguyen, S. Lee and A. C. Bovik, “Dynamic receptive field generation for full-reference image quality assessment,” *IEEE Trans. Image Process.*, pp. 4219-4231, 2020.

[16] Z. Wan, K. Gu and D. Zhao, “Reduced Reference Stereoscopic Image Quality Assessment Using Sparse Representation and Natural Scene Statistics,” *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2024-2037, 2020.

[17] S. A. Golestaneh and L. J. Karam, “Reduced-reference quality assessment based on the entropy of DWT coefficients of locally weighted gradient magnitudes,” *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5293-5303, 2016.

[18] A. K. Moorthy and A. C. Bovik, “Blind image quality assessment: From natural scene statistics to perceptual quality,” *IEEE Trans. Image Process.*, pp. 3350-3364, 2011.

[19] X. Yang, F. Li, W. Zhang and L. He, “Blind image quality assessment of natural scenes based on entropy differences in the DCT domain,” *Entropy*, vol. 20, no. 12, pp. 885-906, 2018.

[20] M. A. Saad, A. C. Bovik, and C. Charrier, “Blind image quality assessment: A natural scene statistics approach in the DCT domain,” *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339-3352, 2012.

[21] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695-4708, 2012.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” *in IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[23] D. Chen, Y. Wang and W. Gao, “No-reference image quality assessment: an attention driven approach,” *IEEE Trans. Image Process.*, vol. 29, pp.6496-6506, 2020.

[24] J. Kim, A. Nguyen, and S. Lee, “Deep CNN-based blind image quality predictor,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 11-24, 2019.

[25] J. Ma, J. Wu, L. Li, W. Dong and X. Xie, "Active inference of GAN for No-Reference Image Quality Assessment," *in Proc. ICME*, pp. 1-6, 2020.

[26] F. Li, Y. Zhang and P. C. Cosman, "MMMNet: an End-to-End Multi-task Deep Convolution Neural Network with Multi-scale and Multi-hierarchy Fusion for Blind Image Quality Assessment," *IEEE Trans. Circuits and Systems for Video Technology*, 2021. (DOI: 10.1109/TCSVT.2021.3055197)

[27] W. Zhang, K. Ma, J. Yan, D. Deng and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36-47, 2020.

[28] X. Yang, F. Li and H. Liu, "Deep feature importance awareness based no-reference image quality prediction," *Neurocomputing*, vol. 401, pp. 209-223, 2020.

[29] K. Ma, W. Liu, Z. Duanmu, Z. Wang and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202-1213, 2018.

[30] Q. Yan, D. Gong, and Y. Zhang, "Two-stream convolutional networks for blind image quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2200-2211, 2019.

[31] X. Yang, F. Li, and H. Liu, "A measurement for distortion induced saliency variation in natural images," *IEEE Transactions on Instrumentation and Measurement*, 2021, (DOI:10.1109/TIM.2021.3108538).

[32] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440-3451, 2006.

[33] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, pp. 19-21, 2010.

[34] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," *in Proc. Asilomar Conf. Signals, Syst. Comput.*, pp. 1693-1697, 2012.

[35] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372-387, 2016.

[36] L. Leveque, J. Yang, X. Yang, P. Guo, K. Dasalla, L. Li, Y. Wu and H. Liu, "CUID: A new study of perceived image quality and its subjective assessment," *in Proc. ICIP*, 2020.

[37] W. S. Geisler, "Visual perception and the statistical properties of natural scenes," *Annu. Rev. Psychol.*, pp. 167-192, 2008.

[38] D. B. Willmore; R. J. Prenger and J. L. Gallant, "Neural representation of natural images in visual area V2," *J. Neurosci.*, pp. 2102-2114, 2010.

[39] Y. Karklin and M. S. Lewicki, "Emergence of complex cell properties by learning to generalize in natural scenes," *Nature*, pp. 83-86, 2009.

[40] M. C. Potter and E. Levy, "Recognition memory for a rapid sequence of pictures," *Journal of Experimental Psychology*, pp. 10573-10581, 2009.

[41] D. B. Walther, E. Caddigan, F. Li and D.M. Beck, "Natural scene categories revealed in distributed patterns of activity in the human brain," *J. Exp. Psychol.*, pp. 10-15, 1969.

[42] F. Li, R. Vanrullen, C. Koch and P. Perona, "Why does natural scene categorization require little attention? Exploring attentional requirements for natural and synthetic stimuli," *Vis. Cogn.*, pp. 893-924, 2005.

[43] F. Li, A. Iyer, C. Koch and P. Perona, "What do we perceive in a glance of a real-world scene?," *Vis. Cogn.*, pp. 11-29, 2007.

[44] B. Tversky and K. Hemenway, "Categories of environmental scenes," *Cogn. Psychol.*, pp. 121-149, 1983.

[45] E. Siahaan, A. Hanjalic, J. A. Redi, "Semantic-aware blind image quality assessment," *Signal Process. Image Commun.* , pp. 237-252, 2018.

[46] H. R. Sheikh, M. F. Sabir and A. C. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440-3451, Nov. 2006.

[47] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal processing magazine*, pp. 130–141, 2017.

[48] X, Yang, F. Li and H. Liu, "A study of DNN methods for blind image quality assessment," *IEEE Access*, pp. 123788-123806, 2019.

[49] Y. Kao, R. He and K. Huang, "Deep aesthetic quality assessment with semantic information," *IEEE Transactions on Image Processing*, pp. 1482-1495, 2017.

[50] Y. Deng, C. Loy and X. Tang, "Image aesthetic assessment: an experimental survey," *IEEE Signal processing magazine*, pp. 80–106, 2017.

[51] K. Ma, W. Liu, K. Zhang, Z. Wang, W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, pp. 1202-1213, 2018.

[52] B. Yan, B. Bare and W. Tan, "Naturalness-aware deep no-reference image quality assessment," *IEEE Transactions on Multimedia*, pp. 2603-2615, 2019.

[53] W. Albert and T. Tullis, "Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics," *Morgan Kaufmann*, 2008.

[54] "Unsplash – Free High-Resolution Photos," [Online]. Available at: https://unsplash.com/.

[55] H. Liu, N. Klomp and I. Heynderickx, "A no-reference metric for perceived ringing artifacts in images," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 529-539, 2010.

[56] S. Athar et al., "Perceptual Quality Assessment of UHD-HDR-WCG Videos," *in Proc. ICIP*, pp. 1740-1744, 2019.

[57] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," [Online]. Available at: http://www.vqeg.org/ (last accessed February 2020).

[58] X. Tang, W. Luo and X. Wang, "Content-based photo quality assessment," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1930-1943, 2013.

[59] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," *in Proc. CVPR*, pp. 770-778, 2016.

[60] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, pp. 1345-1359, 2010.

[61] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *in Proc. ECCV*, 2014.

[62] S. Wang, K. Gu, X. Zhang, W. Lin, L. Zhang, S. Ma, and W. Gao, "Subjective and objective quality assessment of compressed screen content images," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 6, no. 4, pp. 532–543, Dec. 2016.

[63] P. Mohammadi, A. Ebrahimi-Moghadam, and S. Shirani, "Subjective and objective quality assessment of image: A survey," *Majlesi J. Elect. Eng.,* vol. 9, no. 1, pp. 55–83, Mar. 2015.

[64] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *in Proc. ACM MM*, pp. 675-678, 2014.

[65] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," *in Proc. CVPR*, pp. 1733–1740, 2014.

[66] X, Yang, F. Li, and H. Liu, "A survey of DNN methods for blind image quality assessment," *IEEE Access*, vol. 7, no. 1, pp. 123788-123806, 2019.

[67] Z. Wang and A. C. Bovik, "Mean squared error: love it or leave it? - A new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, pp. 98-117, 2009.

[68] H. Sheikh, and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, pp. 430-444, 2006.

[69] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206-220, 2017.

[70] X. Liu, J. van de Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," *in Proc. ICCV*, pp. 1040-1049, 2017.

[71] S. Bosse, D. Maniry, K. Muller, T. Wiegand and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206-219, 2018.

[72] S. Bosse, D. Maniry, K. Muller, T. Wiegand and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206-219, 2018.

[73] S. Sun, T. Yu, J. Xu, W. Zhou and Z. Chen, "GraphIQA: Learning distortion graph representations for blind image quality assessment," 2021, arXiv:2103.07666. [Online]. Available:http://arxiv.org/abs/2103.07666.

[74] X. Yang, F. Li and H. Liu, "TTL-IQA: Transitive transfer learning based no-reference image quality assessment," *IEEE Transactions on Multimedia*, pp. 4326-4340, 2020.

[75] A. Krizhevsky, I. Sutskever, and H. E. Hinton, "Imagenet classification with deep convolutional neural networks," *in Proc. NIPS*, pp. 1097-1105, 2012.

[76] G. Keren, *Between or within-subjects design: A methodological dilemma*, A Handbook for Data Analysis in the Behavioral Sciences, pp. 257-272, 1993.

[77] D. Ghadiyaram and A. C. Bovik, "Massive online crowd-sourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.

[78] Z. Zhu, H. Liu, J. Lu and S. M. Hu, "A Metric for Video Blending Quality Assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 3014-3022, 2020.

[79] Z. Zhu et al., "A Comparative Study of Algorithms for Realtime Panoramic Video Blending," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2952-2965, 2018.

[80] P. Guo, L. He, S. Liu, D. Zeng and H. Liu, “Underwater Image Quality Assessment: Subjective and Objective Methods,” *IEEE Transactions on Multimedia*, 2021. (DOI:10.1109/TMM.2021.3074825)

[81] S. Li and X. Kang, “Fast multi-exposure image fusion with median filter and recursive filter,” *IEEE Trans. Consum. Electron.*, vol. 58, no. 2, pp. 626–632, 2012.

[82] B. Gu, W. Li, J. Wong, M. Zhu, and M. Wang, “Gradient field multi-exposure images fusion for high dynamic range image visualization,” *J. Vis. Commun. Image Represent.*, vol. 23, no. 4, pp. 604–610, 2012.

[83] K. Ma, K. Zeng and Z. Wang, “Perceptual quality assessment for multi-exposure image fusion,” *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, 2015.